*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6477–6485
Marseille, 11–16 May 2020

# CEASR: A Corpus for Evaluating Automatic Speech Recognition

**Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, Mark Cieliebak**
Zurich University of Applied Sciences, Winterthur, Switzerland,
SpinningBytes AG, Winterthur, Switzerland
{ulas, hueu, gedi, benf, ciel}@zhaw.ch, fabian@germann.cc

**Abstract**

In this paper, we present CEASR, a Corpus for Evaluating the quality of Automatic Speech Recognition (ASR). It is a data set based on public speech corpora, containing metadata along with transcripts generated by several modern state-of-the-art ASR systems. CEASR provides this data in a unified structure, consistent across all corpora and systems, with normalised transcript texts and metadata.

We use CEASR to evaluate the quality of ASR systems by calculating an average Word Error Rate (WER) per corpus, per system and per corpus-system pair. Our experiments show a substantial difference in accuracy between commercial versus open-source ASR tools as well as differences up to a factor ten for single systems on different corpora. Using CEASR allowed us to very efficiently and easily obtain these results. Our corpus enables researchers to perform ASR-related evaluations and various in-depth analyses with noticeably reduced effort, i.e. without the need to collect, process and transcribe the speech data themselves.

**Keywords:** automatic speech recognition, evaluation, speech corpus, ASR systems

## 1. Introduction

Automatic Speech Recognition (ASR) has seen rapid progress over the past decades, fuelled by increasing demand for the applications it enables, such as in-meeting assistants, voice search, subtitling, virtual speech assistants, intelligent devices, or dictation tools. Closely linked with the development of high-quality ASR systems is the task of ASR quality evaluation, which is essential for both research and practical applications. The standard metric for measuring ASR performance is the Word Error Rate (WER), which counts the word-level insertions, deletions and substitutions in a generated transcript (*hypothesis*) as compared to a gold standard transcription (*reference*).

For commercial ASR providers, it is common practice to advertise a single WER figure for their system, without mentioning the properties of the audio and speech on which this score was obtained. Similarly, open source models are often evaluated on a very small set of reference corpora, which gives a limited insight into their capability. The *Corpus for Evaluating Automatic Speech Recognition (CEASR*[1]*)* fills this gap by providing transcriptions of nine English and six German speech corpora that were generated by seven different ASR systems. The current version of CEASR (1.0) is a snapshot of state-of-the-art ASR technology from 2019. It can serve as a benchmark for assessing and comparing the quality of various ASR systems as well as tracking the progress in ASR technology over time.[2]

The speech corpora selected for CEASR are standard corpora often cited in the literature. They represent a variety of speaking styles (read-aloud vs. spontaneous, monologue vs. dialogue), speaker demographics (native vs. non-native, different dialectal regions, age, gender and native language), recording environments and audio quality types (e.g. recording studio and telephone line), and thus allow for a nuanced evaluation of ASR performance. The ASR systems reflect the current market and development landscape and as such include both commercial providers and open-source frameworks. To our knowledge, CEASR is the first corpus where transcriptions of multiple ASR systems are collected and published. This allows researchers to explore the capabilities of ASR systems in various settings without the tedious and time-consuming effort of creating the transcriptions. The corpus contains 55'216 utterances with a total of 79'369 tokens from 69.2 hours of audio recordings. References and hypotheses are provided in a unified format, which facilitates the development of scripts and tools for their processing and analysis. CEASR can be used for many applications, including but not limited to: a reference benchmark for ASR quality tracking; detailed error analysis (e.g. error typology by acoustic setting, development of alternative error metrics); detailed ASR quality evaluation (e.g. in relation to speaker demographic profiles or spoken language properties); and improving ASR, for example by developing ensemble learning methods based on the output of different systems.

CEASR has been created from nine English speech corpora (TIMIT, VoxForge, CommonVoice, LibriSpeech, ST, RT-09, AMI, Switchboard 1 Release 2, TedLium) and six German corpora (VoxForge, CommonVoice, Tuda-De, Hempel, StrangeCorpus 10, Verbmobil II v.21). Section 3.1 describes the corpora in more detail. Seven systems, including both commercial and open-source, were used to transcribe the English corpora, while four commercial systems were applied to the German audio recordings.

The remainder of this paper is structured as follows: Section 2 discusses relevant literature, Section 3 describes how CEASR was created, including information on the underlying speech corpora, the ASR systems, as well as the corpus statistics and metadata. In Section 4 we present a detailed analysis of ASR quality as a showcase application of CEASR, before concluding in Section 5.

[1] Pronounced like "Caesar".

[2] The corpus is publicly available at `https://ceasr-corpus.github.io`.

## 2. Related Work

The accuracy of ASR systems has been the subject of numerous investigations. There are several research papers documenting comparative evaluations of ASR systems quality.

Këpuska and Bohouta (2017) show a comparative evaluation of three ASR tools (Microsoft Azure, Google Cloud API, and Sphinx-4) on selected utterances from the TIMIT and International Telecommunication Union (ITU) corpora. Gaida et al. (2014) present a large-scale evaluation of several open-source speech recognition toolkits: HTK with the decoders HDecode and Julius, CMUSphinx with pocketsphinx and Sphinx4, as well as Kaldi are compared in terms of usability and recognition accuracy. The evaluation basis is the Verbmobil 1 (VM1) corpus with dialogues in three languages (English, Japanese and German) as well as the Wall Street Journal 1 (WSJ1) corpus with English read-aloud speech based on Wall Street Journal news. Franck Dernoncourt (2018) developped an ASR benchmark framework which allows users to evaluate seven ASR systems on different corpora. The framework supports the following APIs: Google Chrome Speech API, Google Cloud Speech API, Houndify API, IBM Speech-to-Text, Microsoft Bing Speech-to-Text, Speechmatics, and Wit.ai. The authors also provide scripts allowing to format CommonVoice and LibriSpeech for use in their evaluation framework. Moore et al. (2019) present a meta-dataset for speech intelligibility estimation. This dataset includes the reference text, the hypotheses of two different ASR systems, the number of confusion pairs and total words, the WER, and the time that it took to get the prediction from the ASR system. The corpus is a collection of healthy native, healthy accented, and disordered native speech datasets based on CommonVoice, UASpeech, TORGO, Speech Accent Archive and TIMIT. The hypotheses are generated by Google Speech API and CMUSphinx. This dataset, similarly to CEASR, allows detailed error analysis of the transcripts; however it is much smaller in scope. To the best of our knowledge, CEASR is the most large-scale resource, containing a larger number of corpora and ASR systems than previous work.

## 3. Corpus Description

In the following sections we present CEASR in more detail. Section 3.1 describes the speech corpora that form the basis of CEASR and Section 3.2 details the ASR systems and their configurations. The process of generating CEASR corpus is presented in Section 3.3. The last Section 3.4 is devoted to legal constraints related to the distribution of CEASR.

CEASR consists of a total of 55’216 utterances: 41’527 for English and 13’689 for German. They stem from 15 speech corpora: each sample contains a reference which comes from the underlying corpus and was generated by humans, and a set of machine transcriptions of the corresponding audio recording generated by the ASR systems (hypotheses). The corpus also contains varying metadata, depending on the speech corpus which the utterance comes from. Table 3 contains a detailed description of the metadata.

All utterances in CEASR are examples of either dialogue or monologue human-to-human communication. They exhibit diverse properties related to speaking style, speaking rate, utterance duration, speaker demography or speech disfluencies. The diversity of properties reflect the variability of spoken language. All audio samples have been transcribed with commercial and open-source ASR systems. As a result, each English utterance is provided with between 15 to 19 hypotheses generated by seven commercial and open-source ASR systems with different configurations, and each German sample has four to eight hypotheses by commercial ASR systems.

### 3.1. Speech Corpora

CEASR utterances are derived from nine English and six German public speech corpora. Table 1 provides an overview of the main corpora properties.

The English subset contains three corpora of spontaneous dialogue speech: RT (NIST Multimodal Information Group, 2009), AMI (Carletta, 2006), and Switchboard (Godfrey, John J., and Holliman, Edward, 1997); one consisting of semi-spontaneous monologue utterances: TedLium (Hernandez et al., 2018); and five containing read-aloud monologue speech: Timit (Garofolo et al., 1993), ST (Surfing Technology Ltd, 2018), LibriSpeech Clean and LibriSpeech Other (Panayotov, Vassil and Chen, Guoguo and Povey, Daniel and Khudanpur, Sanjeev, 2015), VoxForge (MacLean, Ken, 2019), and CommonVoice (Mozilla Foundation, 2017). RT and AMI are included in CEASR in two variants: the first one consists of recordings and transcripts from individual headsets, which means the recording contains only the speech produced by one particular speaker. The second version is based on the signals recorded by all the speakers’ head microphones added together. As a result, the recording contains overlapping utterances and noise coming from other speakers. The default test sets were used for TedLium, Timit, VoxForge, CommonVoice and LibriSpeech. Since AMI, RT, and ST do not have standard train/test splits, we created test sets by randomly sampling three hours from RT and ST and five hours from AMI, in proportion to the original corpus sizes. We also took a sample of 5 hours from Switchboard-1 Release 2. This data set is commonly used as training data for telephony scenarios, and therefore it should not be used to evaluate transcription results. Thus, it will not be included in our evaluations in Section 4, but as it is suitable for other ASR research purposes, it is included in CEASR.

The German part of CEASR contains transcriptions of four corpora with read-aloud monologue utterances: CommonVoice (Mozilla Foundation, 2017), VoxForge (MacLean, Ken, 2019), Tuda-De (Milde and Koehn, 2015), and the Strange Corpus 10 subset with read-aloud speech (Mapelli, 2004); two corpora containing spontaneous monologue: Hempel (Draxler, 2004) and the Strange Corpus 10 subset with retelling speech (Mapelli, 2004); and two with spontaneous dialogue speech: Verbmobil II v.21 (BMBF, Projekträger DLR, 2004) and the Strange Corpus 10 subset with spontaneous speech (Mapelli, 2004). Two German corpora (Tuda-De and CommonVoice) are provided with default test sets. For all other corpora, random samples of

| | **English** | **German** |
|---|---|---|
| Speaking styles | Read-aloud monologue, semi-spontaneous monologue, spontaneous dialogue | Read-aloud monologue, spontaneous monologue, spontaneous dialogue |
| Number of corpora including corpora subsets and variants | 13 (including two subsets of LibriSpeech, two variants of RT, AMI and TedLium) | 10 (including three subsets of Strange Corpus 10) |
| Size in hours | 47.55 | 21.63 |
| Number of utterances | 41'527 | 13'689 |
| Number of tokens | 50'777 | 28'592 |
| Average utterance duration (in seconds) | 4.61 sec | 7.67 sec |
| Number of identified speakers | 1'351 | 723 |
| Gender distribution | 781 male and 509 female speakers | 369 male and 364 female speakers |
| Number of identified non-native speakers | 148 | 194 |
| Native languages of identified non-native speakers | n/a | en-US, pt, it, es, ar, tr-TR, en-GB, ru-RU, nl, de-AT, ja-JP, fi-FI, el-GR, de-CH, fr, pl-PL, de, sv, hu-HU |
| Mean, minimal and maximal number utterances per speaker | mean: 107, min: 1 (Voxforge and Switchboard), max: 757 (RT both variants) | mean: 32, min: 1 (Voxforge and Hempel), max: 832 (VoxForge) |
| Number of utterances per dialect | en-US: 15'200, en-GB: 1044, en-CA: 508, en-AU: 176, en-IE: 162, en-IN: 150, en-NZ: 79, en-PH: 8, en-ZA: 7, en-MY: 3, en-SG: 1, unknown: 24'189 | de-DE: 4869, de-CH: 84, de-AT: 47, unknown: 8'689 |
| Average speaking rate | 148 words / min | 140 words / min |
| Number of utterances containing filled pauses | 10'992 | 1'218 |
| Number of utterances containing only filled pauses | 4'331 | 0 |
| Number of utterances containing only speaker noise | 1'380 | 0 |
| Distribution of overlapping utterances (only spontaneous dialogue speech corpora) | 78.8 % (RT both variants), 84.4 % (AMI both variants), 90.5 (Switchboard) % | n/a |

Table 1: Properties of standard speech corpora forming the basis of CEASR.

three hours were taken, except for the subsets of Strange Corpus 10: due to their small volumes, the complete sets of utterances were used.

## 3.2. ASR Systems

Some commercial systems have confidentiality restrictions in their Terms of Use, so we cannot disclose their names or associate them with the WER scores they obtained. There is one exception: Microsoft Azure Speech-to-Text[3] gave permissions to be named and is labelled as System 6. In the remainder of this paper, we will refer to the other commercial providers simply as Systems 1, 2 and 3.

The open-source systems have no such constraints, so we can provide their names, but for unity of format we will also use numbers to refer to them in the graphics: System 4 is Kaldi version 5.5, System 5 is Mozilla DeepSpeech version 0.5.1 and System 7 is CMUSphinx sphinx4.

The English corpora were transcribed by both commercial and open-source systems, while the German part of CEASR 1.0 was only transcribed with commercial systems. The integration of open-source systems for German is currently ongoing and will be part of a future release of CEASR.

The systems differ in terms of applied ASR paradigms: CMUSphinx is based on Hidden Markov Models (HMMs) (Dhankar, 2017), while Mozilla implements a more recent End-to-End architecture, where the entire ASR process is performed with a single Neural Network ((Amodei et al., 2016), (Hannun et al., 2014)). Kaldi (Povey et al., 2011) represents the hybrid approach combining components using HMMs, Gaussian Mixture Models (GMMs) and Deep Neural Networks (DNNs).

We have used the cloud (online) versions of all the commercial providers and local (offline) installations of the open-source systems. However, Systems 1, 3 and Microsoft also have on-premise options, which we have not yet investigated.

All the systems have been used with their standard models without any customisation or additional training. We note that this might influence results as some systems have been optimised for specific use cases or require customisation for best results. Mozilla DeepSpeech and Kaldi are the only systems where we have information about the speech corpora used for default model training. Mozilla DeepSpeech was trained on the Fisher corpus (Cieri et al., 2004), LibriSpeech and Switchboard; Kaldi was trained on Fisher (model ASpIRE) and LibriSpeech (model LibriSpeech). We also know that S§ystem 2 has not been trained on any of the corpora integrated in CEASR. For the remaining systems, we do not know at this point whether their training sets overlap with the corpora from CEASR.

We performed transcriptions for each utterance with various settings per system, trying to find the configurations with the highest transcription accuracy. Table 3.2 provides an overview of the configurations of the three open-source systems.

| **System** | **Model** | **Audio Properties** | **Language** |
|---|---|---|---|
| *Mozilla DeepSpeech* | *Included* | *WAV / 16 kHz / 16 bit* | *en-US* |
| *Kaldi* | *LibriSpeech* | *WAV / 16 kHz / 16 bit* | *en* |
| Kaldi | LibriSpeech | WAV / 8 kHz / 8 bit | en |
| Kaldi | ASpIRE | WAV / 16 kHz / 16 bit | en |
| Kaldi | ASpIRE | WAV / 8 kHz / 16 bit | en |
| CMUSphinx | Included, PTM | WAV / 16 kHz / 16 bit | en-US |
| *CMUSphinx* | *Included, Continuous* | *WAV / 16 kHz / 16 bit* | *en-US* |

Table 2: Configurations of open-source systems. Italics mark best configuration for each system.

For the four commercial systems, the number of available configuration options varied significantly. For two out of four systems, more than one model was available. Apart from model selection, the commercial systems also allowed

[3] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

setting different audio sampling rates and bit depths. We experimented with two different sampling rates (8 kHz and 16 kHz) and bit depth configurations (8 bit and 16 bit). Some systems allowed setting additional parameters related for example to transcripts formatting. We cannot disclose more details on the configuration of the anonymous commercial systems to ensure their anonymity.

### 3.3. Corpus Generation Process

The process for generating CEASR consisted of the following steps: pre-processing of the speech corpora in order to extract and normalise utterances; transcribing the utterances with the ASR systems; and post-processing the generated transcripts. The steps are described below.

**Corpus extraction** After retrieving the data from each corpus, we performed a quality check in order to identify incomplete utterance data: if the original reference text or the audio file were missing, or if the utterance start time was after or equal to the utterance end time, we removed the utterance from the data set. This step was necessary to ensure all data required for generating a transcription and aligning it with its reference was available. In total, we removed 549 utterances from the German corpora. Next, we normalised the metadata items gender, dialect, and mother tongue in order to allow cross-corpus filtering and comparison of utterances. For dialect and mother tongue, the original data was mapped to the ISO 639-1 standard language codes with country codes (e.g. en-US, de-AT, etc.); for gender, all entries were normalised to 'male' and 'female'. Finally, we transformed each corpus into a uniform utterance data structure needed for further processing. Table 3 provides more details on the structure of the utterance data object. More detailed documentation is provided on the corpus website.

**Reference segmentation and normalisation** If the references were provided in one large file containing multiple annotated utterances, we retrieved each single utterance and stored it separately according to the annotation. We aimed at keeping the utterance lengths consistent and not exceeding ten seconds by segmenting longer utterances. However, we could perform this segmentation only for utterances provided with time stamps. If no time stamps were available, the utterances were left as provided in the original corpus. As a result, mean utterance duration for English ranges from 2.03 (RT Headset) to 8.16 seconds (TedLium) and for German lies between 2.78 (Strange Corpus 10 read-aloud subset) and 24.79 seconds (Hempel).

Furthermore, we removed all meta-tags from the reference text. These are tags describing speaker noise (e.g. laughing or coughing), non-speaker noise (e.g. rustle, squeak), or speaker disfluencies (e.g. speaker restarts, partial words, mispronounced words, unintelligible speech). Next to the cleaned reference, the original reference text including the meta-tags was stored as part of the unified utterance data as well. A full documentation of meta-tags is available on the CEASR website.

In order to reliably compare performance between different systems, the formatting of the references and the hypotheses must be as similar as possible. To this end, we removed punctuation, transformed all strings to lower-case, spelled out numbers (”forty-two” instead of ”42”), and applied consistent formatting of integers, decimal values and time of day.

**Audio pre-processing** In order to prepare utterances for transcription, we performed two steps: extracting audio information, converting and trimming the audio file where necessary.

Some speech corpora provide a set of audios of several seconds, each containing one short utterance, while others have long audios with long utterances of several minutes, or long audios containing multiple short utterances. When timestamps were available, we segmented all recordings into audios of duration less than 10 seconds in order to ensure consistency between reference text and audio segmentation (see previous paragraph). In the next step, we converted the audios into the required audio formats. We also changed the sampling rate and the bit depth according to the requirements by the particular system. Converted audios were then passed through to the transcription step and the utterance data object was extended with new information such as format, duration, sampling rate, bit depth and number of channels of both the original audio file as well as the converted recording.

**Corpus transcription** The transcriptions were performed in batches: a full set of pre-processed utterances from one corpus was sent to an ASR system for transcribing. Each utterance data object was enriched with the original hypothesis text and moved to the next step, which was hypothesis post-processing

**Hypothesis post-processing** In order to ensure consistency between reference and hypothesis, we performed the same reference normalisation steps also on the hypothesis(see paragraph ”Reference segmentation and normalisation”). We tried to discover and eliminate as many discrepancies between references and hypothesis as possible; however, a complete consistency between the texts in terms of formatting cannot be guaranteed.

As a result, CEASR contains utterances of the speech corpora together with their metadata, the transcripts and the transcription details stored in a unified format. As shown in Table 3, it covers various speaking styles; a variety of speakers' demographic profiles related to gender, dialect and mother tongue; various recording setups (microphone types and recording environments); properties of the audio signal (duration, sampling rate, number of channels, bit depth, encoding and number of samples), as well as other spoken utterance properties such as speaking rate, occurrence of speaker noise (e.g. laughing) or filled pauses (such as e.g. ”hm” or ”mhm”). Utterances in the corpus are provided with a set of attributes reflecting all these dimensions. The utterances from one speech corpus transcribed by one system in a particular configuration are stored in one file, which also contains job metadata: the corpus name and system configuration as well as metadata such as language, speaking style, original reference and audio segmentation, and dialogue or monologue speech categorisation.[4]

[4] More detailed corpus documentation is provided on the CEASR website.

| Utterance Attribute | Description |
|---|---|
| Utterance ID | Utterance ID (unique within corpus). |
| Speaker ID | Identifier of the speaker if available. |
| Reference | The manual transcript provided as part of a speech corpus. Two variants of the reference are stored: original transcript and processed transcript (according to the pre-processing steps described in Section 3.3) |
| Hypothesis | The machine transcript generated by an ASR system and post-processed according to the steps described in Section 3.3. The hypothesis section also contains the original machine transcript generated by the ASR system. |
| Audio | Detailed information about the audio recording of the utterance: audio file name, audio duration, sampling rate, bit depth, number of channels, encoding and number of samples. |
| Recording | Type of the recording device and acoustic environment (where available). |
| Dialect | The ISO 639-1 language code with country code describing the speaker's dialect. |
| Accent | The accent of the speaker (native or non-native) |
| Gender | The gender of the speaker. The value is normalised across all corpora and has the value 'male' or 'female'. |
| Overlappings | Whether the utterance is overlapping with any other utterance. |
| Speaker noise | Whether the utterance contains only speaker noise such as laughter. |
| Additional properties | Additional utterance properties, such as mother tongue of a non-native speaker, region, education or age. (Available for a limited number of utterances). |
| Speaking rate | The number of words uttered per minute. |

Table 3: Utterance data object.

### 3.4. Constraints on CEASR Distribution

Due to legal constraints, not all of the corpus references and system hypotheses can be made publicly available. The information below reflects the current status at the time of writing; however, since this is subject to change, a detailed overview of which utterances and transcripts are publicly available as part of the latest CEASR version will be documented on the CEASR website.

**Corpora constraints** References from corpora with Creative Commons or GNU General Public licenses (AMI, CommonVoice, LibriSpeech, ST, Tedlium and VoxForge, Tuda-De) can be shared without restrictions. References from the remaining corpora are not published as part of CEASR 1.0. Details on how to integrate utterances from paid corpora are provided on the CEASR website.

**System constraints** Some providers place restrictions on distributing the hypotheses generated by their systems. Transcripts by the open-source systems (Mozilla, Sphinx4 and Kaldi) can be distributed. Furthermore, three commercial systems have given permission for transcript distribution and are included in CEASR 1.0. The names of these commercial systems, however, cannot be disclosed.

## 4. Sample Application: Using CEASR for ASR System Comparison

CEASR lends itself to a detailed evaluation and comparison of ASR systems. In the following, we first briefly discuss what it means to run a fair evaluation (Section 4.1), explain what corpora form the basis of our evaluation (Section 4.2) and finally we evaluate the quality of ASR systems for English and German (Section 4.3). We analyse the performance with respect to the characteristics of corpora (e.g. speaking style - see Sections 4.4 and 4.6, or accent - see Section 4.7), and commercial versus open-source systems (4.5).

### 4.1. Challenges for a Fair Evaluation

Before we discuss the evaluation results, we would like to point out some of the caveats of running a fair comparison between the hypotheses of different systems on the one hand and the corpus references on the other hand. In the following we will describe the issues encountered. Although we have completed this analysis to the best of our knowledge, it remains possible that there are hitherto undiscovered biases in the data.

**Treatment of filled pauses** Not all systems output filled pauses such as "uh-huh" or "mhm", which are very frequent in conversational speech, as part of their hypotheses. Investigating utterances from the English spontaneous speech corpora AMI and RT, we found that while Kaldi produces filled pauses, System 2 and 3, Mozilla and Sphinx tend to omit them. For two commercial systems, no clear tendency was detected[5]. Since filled pauses such as "mhm" are not removed from the references, systems that do not produce output for them will have higher WER scores for utterances that include filled pauses.

**Different spelling conventions** The hypotheses of two German corpora, Verbmobil II v.21 and Strange Corpus 10, use German spelling prior to the 2006 spelling reform (e.g. 'daß' instead of 'dass'), which results in somewhat misleading WERs. We aim to normalise such diachronic spelling variants in a future version of CEASR. For now, we ask the reader to bear in mind that the absolute WERs on these corpora cannot be taken at face value, but the difference between individual systems should be represented correctly.

[5] The WERs of the two systems do not change much if we keep filled pauses in the transcripts instead of removing them. The difference between the conditions with/without filled pauses for these two systems was in the range of 1-2% absolute WER, whereas for the other systems it was at least 10%. We have not investigated this further but believe that it is due to inconsistent treatment of filled pauses by these two systems.

**Spelling variants** There are also differences with respect to how systems spell common words. In German, for example, Systems 1 and 3 transcribe *ok* as ”ok”, while all other systems generate ”okay”, which is also the most likely form in the corpus references. Such spelling differences have not been normalised for the evaluation.

**Training data bias** We also ask the reader to bear in mind that for some commercial models, we have no knowledge of the training corpora that were used (see Section 3.2 for details). This causes a potential bias in our results since it is possible that they were trained on some of the evaluation data used in CEASR.

## 4.2. Dataset for System Comparison

We used the same set of utterances for the calculation of results in Sections 4.3, 4.4, 4.5 and 4.6: we selected the best-performing configuration of each system. We also excluded two English corpora: Tedlium Unsegmented to keep utterance lengths consistently short in the English subset, and Switchboard due to possible bias, see Section 3.1.

In the English corpora, we additionally removed utterances with overlapping speech, which removed 78.8% from both RT corpora variants and 84.4% from both AMI datasets. We could not perform overlapping detection on the German corpora due to missing time stamps.

For each of the utterances, we compared the generated hypothesis of each system to the reference, and calculated the WER using sclite[6]. The final metric for each corpus-system pair consists of the average WER across all utterances in the corpus transcribed by the system.

## 4.3. General Comparison of ASR Systems

Bearing in mind the limitations described in Section 4.1, we will now use CEASR to do a detailed comparison of the WER performance of different systems.

Figure 1 presents an overview of the results obtained for English. Each cell in the heatmap represents a corpus-system pair, sorted according to average WER (top-down for systems and left to right for corpora). For each system the best performing configuration was selected.

We can see that the top three systems (S2, Microsoft Azure and S1) are performing consistently well. The last commercial system, S3, has substantially higher WERs, especially on spontaneous speech corpora.

Figure 2 shows the same evaluation for German. The top two systems are the same as for English, however with a different ranking: Microsoft Azure in the first place and S2 in the second. Similarly to English, S3 significantly falls behind.

Table 4 shows the aggregated results by speaking style and system type for English. We can see that spontaneous speech has higher WERs than read-aloud speech, and that commercial systems are on average more performant than open-source systems. The distinction between speaking styles can also be seen for German, where the average WER on spontaneous speech is 25.9%, while it is 14.8% for read-aloud speech. We will elaborate these dimensions in more detail below: Section 4.4 discusses the differences in WER due to speaking style and Section 4.5 looks at commercial versus open-source systems for English.

| | Read-aloud and Semi-spontaneous Speech | Spontaneous Speech |
|---|---|---|
| Commercial Systems | 11.6% | 35.5% |
| Open-source Systems | 26.6% | 64.7% |

Table 4: Average WER results on English corpora grouped by speaking style and system type.

## 4.4. Spontaneous Speech has Higher WERs

Figures 1 and 2 show that spontaneous dialogue speech recognition is substantially more challenging for all systems than recognising non-spontaneous or semi-spontaneous speech: the lowest WER obtained on spontaneous English speech are six times larger than the lowest WER on read-aloud speech (RT Headset - 24.9% vs ST - 4.4%) .

For German, the best spontaneous transcription has a WER almost three times larger than the best read-aloud (Hempel - 14.83% vs VoxForge - 6.03%). In general, the discrepancy between speaking styles is not as large as for English: the WERs on the spontaneous monologue corpus Hempel are only between 2.4% and 7.4% absolute higher than the WER on the read-aloud corpus CommonVoice, depending on the system. Verbmobil II is the most challenging corpus on average, but this can be partly explained by its use of outdated German spelling - please refer to Section 4.1 for details.

These results are consistent with the literature: spontaneous speech differs from read-aloud speech in ways which makes it more difficult to recognise, both acoustically and linguistically (Nakamura et al., 2008). Dialogue speech is challenging due to non-canonical pronunciations, acoustic and prosodic variability, and high levels of disfluency, e.g. repetitions, false starts and repaired utterances ((Goldwater et al., 2010), (Hassan et al., 2014)). Spontaneous speech is additionally characterised by accelerated speaking rates and higher proportions of out-of-vocabulary words (Nakamura et al., 2008).

## 4.5. English: Commercial Systems Outperform Open-Source Systems

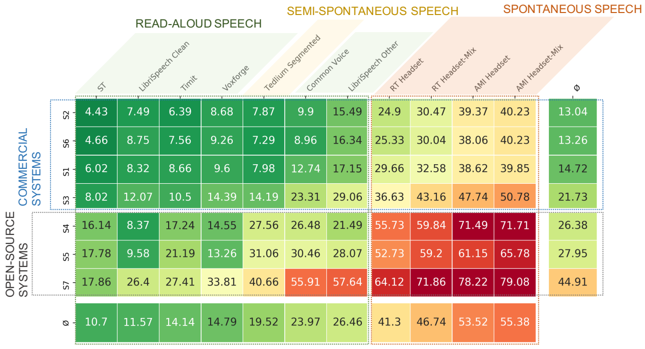
In Figure 1 we can see that the best commercial system obtains 4.43% WER on read-aloud speech (S2 on ST) and 24.9% on spontaneous speech (S2 on RT Headset), which contrasts with the best results obtained by an open-source system: 8.37% on read-aloud (Kaldi on LibriSpeech Clean) and 52.73% on spontaneous speech (Mozilla on RT Headset).

Figure 3 shows the WERs of the best and worst (in terms of global average WER) English commercial and open-source systems, as well as the average across all commercial and open-source systems, respectively. On average, the WERs of commercial systems are lower than those of open-source systems by a factor 2, and the best commercial system performance is unreachable for any open-source system, with

[6] https://github.com/usnistgov/SCTK

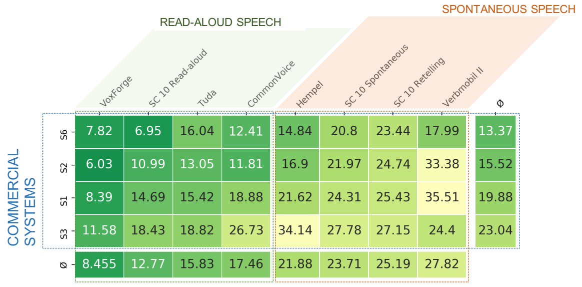
the exception of the LibriSpeech corpora, where Kaldi outperforms S3, and VoxForge, where their performance is the same (remember, however, that the best Kaldi model, Kaldi LibriSpeech, was trained on LibriSpeech).

It is evident from these data that commercial cloud providers currently have an advantage over pre-trained open source solutions. We hypothesise that the big technology companies, which provide the commercial systems, have much larger proprietary data sets at their disposal, resulting in better performance across different speech corpora. It would be interesting to investigate whether using the open-source systems to train more customised models could offset this training data disadvantage.



| | | READ-ALOUD SPEECH | | | SEMI-SPONTANEOUS SPEECH | | | SPONTANEOUS SPEECH | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ST | LibriSpeech Clean | Timit | Voxforge | Tedlium Segmented | Common Voice | LibriSpeech Other | RT Headset | RT Headset-Mix | AMI Headset | AMI Headset-Mix | ϕ |
| COMMERCIAL SYSTEMS | S2 | 4.43 | 7.49 | 6.39 | 8.68 | 7.87 | 9.9 | 15.49 | 24.9 | 30.47 | 39.37 | 40.23 | 13.04 |
| | S6 | 4.66 | 8.75 | 7.56 | 9.26 | 7.29 | 8.96 | 16.34 | 25.33 | 30.04 | 38.06 | 40.23 | 13.26 |
| | S1 | 6.02 | 8.32 | 8.66 | 9.6 | 7.98 | 12.74 | 17.15 | 29.66 | 32.58 | 38.62 | 39.85 | 14.72 |
| | S3 | 8.02 | 12.07 | 10.5 | 14.39 | 14.19 | 23.31 | 29.06 | 36.63 | 43.16 | 47.74 | 50.78 | 21.73 |
| OPEN-SOURCE SYSTEMS | S4 | 16.14 | 8.37 | 17.24 | 14.55 | 27.56 | 26.48 | 21.49 | 55.73 | 59.84 | 71.49 | 71.71 | 26.38 |
| | S5 | 17.78 | 9.58 | 21.19 | 13.26 | 31.06 | 30.46 | 28.07 | 52.73 | 59.2 | 61.15 | 65.78 | 27.95 |
| | S7 | 17.86 | 26.4 | 27.41 | 33.81 | 40.66 | 55.91 | 57.64 | 64.12 | 71.86 | 78.22 | 79.08 | 44.91 |
| | ø | 10.7 | 11.57 | 14.14 | 14.79 | 19.52 | 23.97 | 26.46 | 41.3 | 46.74 | 53.52 | 55.38 | |

Figure 1: General overview of WER results for best performing system configurations for English.



| | | READ-ALOUD SPEECH | | | | SPONTANEOUS SPEECH | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | VoxForge | SC 10 Read-aloud | Tuda | CommonVoice | Hempel | SC 10 Spontaneous | SC 10 Retelling | Verbmobil II | ϕ |
| COMMERCIAL SYSTEMS | S6 | 7.82 | 6.95 | 16.04 | 12.41 | 14.84 | 20.8 | 23.44 | 17.99 | 13.37 |
| | S2 | 6.03 | 10.99 | 13.05 | 11.81 | 16.9 | 21.97 | 24.74 | 33.38 | 15.52 |
| | S1 | 8.39 | 14.69 | 15.42 | 18.88 | 21.62 | 24.31 | 25.43 | 35.51 | 19.88 |
| | S3 | 11.58 | 18.43 | 18.82 | 26.73 | 34.14 | 27.78 | 27.15 | 24.4 | 23.04 |
| | ø | 8.455 | 12.77 | 15.83 | 17.46 | 21.88 | 23.71 | 25.19 | 27.82 | |

Figure 2: General overview of WER results for best performing system configurations for German.

### 4.6. English: WERs of the Same System Differ Drastically Across Corpora

The difference within the same system on various corpora is larger than within the same corpus between systems, as visible in Figure 1. The largest relative difference can be observed for S2, which has a WER of 4.43% on ST and a WER of 40.23% on AMI Headset Mix (a factor 10). When comparing the performance of all systems within these two corpora, it can be observed that the difference is up to a factor 4.2. WER on ST spans between 4.43% (S2) and 17.86% (Sphinx) and on AMI Headset Mix it ranges from 39.85% (S1) to 79.08% (Sphinx).

Figure 4 presents the WERs on the English corpora for the best and the worst performing system per corpus. It can be observed that the read-aloud corpus with the lowest error rate is in both cases the same: ST. The worst score has been achieved on AMI Headset Mix for both the worst performing system and the best performing system. The difference is substantial and can be explained by the properties of the corpora. ST exhibits significantly different properties than AMI Headset Mix, which are much less challenging in terms of ASR: it contains only native speech recorded in a silent indoor environment, it does not contain any utterances with filled pauses and neither with speaker noise, while AMI consists of recordings of non-native speakers in a meeting room with multiple participants, the utterances contain speaker noise and filled pauses. Table 5 compares the key properties of both corpora.

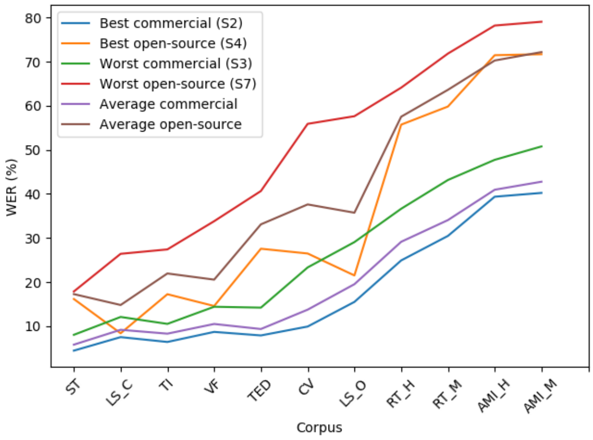


Figure 3: Best and average commercial and open-source systems for English per corpus.

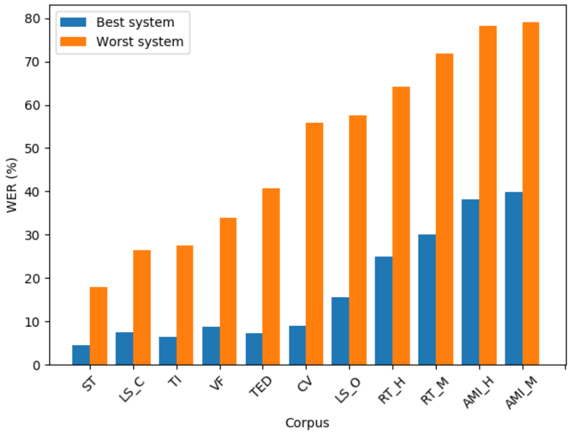


Figure 4: Best and worst scores on English corpora.

### 4.7. English: ASR Quality Substantially Deteriorates on Non-Native Speech

This experiment investigates the impact of native and non-native speaker accents on ASR system performance for English. We selected the subset of English CEASR utterances without overlapping speech for which accent information was provided. We only considered corpora that contained

| | **ST** | **AMI Headset Mix** |
|---|---|---|
| Avg WER on all systems | 10.7% | 55.38% |
| Number of utterances | 2422 | 4495 |
| Average utterance duration | 4.46 | 3.95 |
| Number of speakers | 5 | 190 |
| Number of non-native speakers | 0 | 65 |
| Dialect(s) | en-US | en-AU, en-IE, en-GB, en-CA, en-IN |
| Vocabulary size (no stemming) | 3391 | 3546 |
| Average speaking rate | 109 words/min | 141 words/min |
| Acoustic environment | silent indoor environment | meeting room |
| Recording device | unknown | headset |
| Utterances with filled pauses | 0 | 2160 |

Table 5: Characteristics of corpora with the lowest and the highest average WER.

both native and non-native utterances in order to minimise corpus effects. This set consists of a total of 1629 utterances (1058 by non-native and 571 by native speakers) retrieved from AMI (503 utterances) and RT (1126 utterances). There are 94 distinct speakers in the native subset and 61 in the non-native one.

Figure 5 shows the results: the coloured bars represent the mean of the set, and the dots the results of individual systems (from left to right: Systems 1 to 7). We can see that non-native speech leads to higher WERs than native speech, both on average and for each of the systems individually. The average WER for non-native speech is 7.92% absolute higher than for native speech, with deltas for individual systems ranging from 5% (Sphinx) to 13% (Mozilla).

Some of the main challenges for ASR on non-native speech are disfluencies, accented pronunciation, pronunciation errors due to unfamiliarity with a word, errors in syntax, and syntax that is unusual but not incorrect.

Other difficulties include overemphasis of word boundaries. The irregularity of these deviations makes the speech recognition task even more challenging (Tomokiyo, 2000). Furthermore, ASR systems are usually trained on native speech, which leads to a discrepancy between the WER of native versus non-native speech recognition.

## 5. Conclusion

”We have presented CEASR, a new corpus for evaluating ASR with 55’216 utterances from 15 audio corpora, which amount to 69.2 hours of audio, transcribed by seven systems in different configurations. As a showcase, we have used CEASR to evaluate the quality of state of the art systems for ASR. This comparative analysis showed that the differences in the transcription quality between the systems, but also across the corpora, can be substantial. We observed differences up to a factor of ten between the results within the same ASR system but on different corpora. We have seen that, among the seven systems under investigation, the commercial systems obtained significantly better results than the open-source systems. Based on the results analysis, we could also identify some major challenges that modern ASR systems are facing, especially when applied for meeting or interview transcriptions: spontaneous speech containing disfluencies, speaker and non-speaker noise as well as non-native speech cause a significant increase in WER.

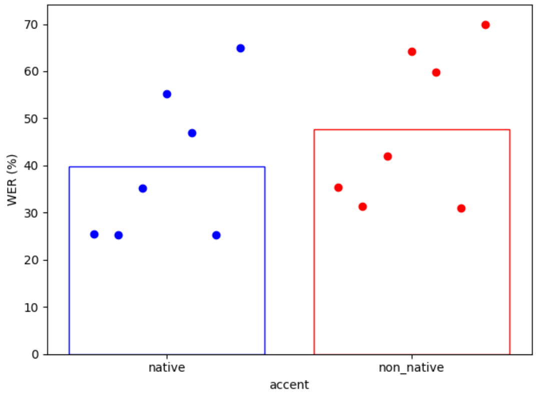


Figure 5: WERs for English native versus non-native speech. Dots represent individual results by Systems 1-7 (left to right).

CEASR allowed to easily perform the evaluation mentioned above, after performing alignment and WER calculation. We then developed scripts for generating statistics and charts, taking advantage of the uniform format for all sub-corpora and systems. CEASR thus provides researchers with a data set ready to be applied in many ASR-related research projects with minimal effort. Transcriptions generated by other systems and further corpora will be added in future versions of CEASR. We also intend to extend it to more languages. The corpus version discussed in this paper will be referenced as CEASR 1.0 in our future publications. Apart from extending the content of the corpus, we also intend to apply it for further research, such as evaluating the semantic relevance of the WER metric, as well as using the CEASR transcripts to explore ensemble methods for improving ASR accuracy.

## 6. Acknowledgements

This research was funded by CTI grant 28190.1 PFES-ES CALL-E. We would like to thank all the ASR system providers and Comparis Schweiz AG for their support. We would furthermore like to thank Lina Scarborough for her helpful comments on the manuscript.

## 7. Bibliographical References

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al. (2016). End to End Speech Recognition in English and Mandarin.

Dhankar, A. (2017). Study of Deep Learning and CMU Sphinx in Automatic Speech Recognition. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2296–2301. IEEE.

Franck Dernoncourt, Trung Bui, W. C. (2018). A Framework for Speech Recognition Benchmarking. In *Interspeech*.

Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., and Suendermann-Oeft, D. (2014). Comparing Open-Source Speech Recognition Toolkits. *Tech. Rep., DHBW Stuttgart*.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep Speech: Scaling up End-to-End Speech Recognition. *arXiv preprint arXiv:1412.5567*.

Hassan, H., Schwartz, L., Hakkani-Tür, D., and Tur, G. (2014). Segmentation and disfluency removal for conversational speech translation. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Këpuska, V. and Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Int. J. Eng. Res. Appl*, 7(03):20–24.

Moore, M., Saxon, M., Venkateswara, H., Berisha, V., and Panchanathan, S. (2019). Say what? A dataset for exploring the error patterns that two ASR engines make. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2019, pages 2528–2532.

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between Acoustic Characteristics of Spontaneous and Read Speech and their Effects on Speech Recognition Performance. *Computer Speech & Language*, 22(2):171–184.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Tomokiyo, L. M. (2000). Handling Non-Native Speech in LVCSR: A Preliminary Study. In *Proceedings of the EUROCALL/CALICO/ISCA Workshop on Integrating Speech Technology in (Language) Learning (InSTIL)*.

## 8. Language Resource References

BMBF, Projektträger DLR. (2004). *VERBMOBIL II - VM CD21.1 - VM21.1 (ELRA-S0034-30)*. European Language Resources (ELRA), 1.0, ISLRN 837-421-490-699-3.

Carletta, Jean. (2006). *The AMI Meeting Corpus*. AMI Consortium.

Cieri, Christopher and Graff, David and Kimball, Owen and Miller, Dave and Walker, Kevin. (2004). *Fisher English Training Speech Part 1 Transcripts*. Linguistic Data Consortium (LDC), ISLRN 100-086-600-941-5.

Draxler, Christoph. (2004). *Hempel (ELRA-S0162)*. European Language Resources (ELRA), 1.0, ISLRN 683-410-635-177-8.

Garofolo, John S., and Lamel, Lori F. and Fisher, William M. and Fiscus, Jonathan G. and Pallett, David S. and Dahlgren, Nancy L. and Zue, Victor. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium (LDC), LDC93S1, ISLRN 664-033-662-630-6.

Godfrey, John J., and Holliman, Edward. (1997). *Switchboard-1 Release 2*. Linguistic Data Consortium (LDC), ISLRN 988-076-156-109-5.

Hernandez, François and Nguyen, Vincent and Ghannay, Sahar and Tomashenko, Natalia and Estève, Yannick. (2018). *TED-LIUM Release 3*.

MacLean, Ken. (2019). *VoxForge*.

Mapelli, Valérie. (2004). *Strange Corpus 10 - SC10 ('Accents II') (ELRA-S0114)*. ELRA (via CLARIN VLO), 1.0, ISLRN 024-991-750-952-3.

Milde, Benjamin and Koehn, Arne. (2015). *Tuda-De: Open Speech Data for German Speech Recognition*. Dialog+ Project.

Mozilla Foundation. (2017). *CommonVoice*. Mozilla Foundation.

NIST Multimodal Information Group. (2009). *Rich Transcription 2009 Meeting Recognition (RT-09) Evaluation Set*.

Panayotov, Vassil and Chen, Guoguo and Povey, Daniel and Khudanpur, Sanjeev. (2015). *Librispeech: An ASR corpus based on public domain audio books*. Institute of Electrical and Electronics Engineers (IEEE).

Surfing Technology Ltd. (2018). *ST-AEDS-20180100_1, Free ST American English Corpus*. Open Speech and Language Resources (OpenSLR).